Automated Classification of Osteoarthritis in the Knee Using Deep Learning

Simon Olsson

Karolinska Institutet

Ehsan Akbarian

Karolinska Institutet

Anna Lind

Karolinska Institutet

Ali Sharif Razavian

Karolinska Institutet

Max Gordon (✉ max@gforge.se)

Karolinska Institutet, Stockholm, Sweden <https://orcid.org/0000-0002-8080-5815>

Research article

Keywords: Deep learning, artificial intelligence, knee osteoarthritis, radiographs, Kellgren & Lawrence.

Posted Date: January 29th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-154385/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

BACKGROUND

Prevalence for knee osteoarthritis is rising in Sweden and globally due to an ageing and more obese population. This has subsequently led to an increasing demand for knee arthroplasties. Correctly diagnosing, classifying, follow-up and planning for either conservative or operative management of knee OA is therefore of a great interest. Most orthopedic surgeons rely on standard weight bearing radiographs, improving the reliability and reproducibility of these interpretations could thus be hugely beneficial. Recently deep learning, a form of artificial intelligence (AI), has shown promising results for interpreting radiographs. In this study, we aim to evaluate how well an AI can classify knee OA severity using entire image series and not excluding common visual disturbances such as implants, casts and other pathologies.

Methods

We selected 6103 radiograph exams taken at Danderyd University Hospital between the years 2002-2016 and manually categorized them according to the Kellgren & Lawrence grading scale (KL). We then trained a convolutional neural network that we evaluated against a test set of 300 exams. These exams had been reviewed independently by two senior orthopedic surgeons who settled exams with disagreement through a consensus session.

Results

Our network yielded an overall high AUC of >0.87 for all KL grades except KL grade 2 and a mean AUC of 0.92. When merging adjacent KL grades together, all but one group showed near perfect results with AUC > 0.95 indicating excellent performance.

Conclusion

We found that we could teach a neural network classify knee OA severity and laterality using the KL grading scale without cleaning the input data from major visual disturbances such as implants and other pathologies.

Background

With an aging population and increasing obesity worldwide, the prevalence of knee osteoarthritis (OA) is higher compared to other types of OA (1). Knee arthroplasties in Sweden is expected to increase considerably the coming decade (2). Similarly in the US, half of the population may have developed knee OA by the age 85 (3) and the demand for total knee arthroplasty (TKR) is expected to be 6 times more common in 2030 as compared to 2005 (4). Correctly diagnosing, classifying, follow-up and planning for either conservative or operative management of knee OA is therefore of a great interest.

The diagnostic criteria of knee OA consist of a combination of pain, clinical and radiological findings. While pain is the key, it is highly elusive and difficult to quantify (5). Radiographs in turn, correlate with number of symptoms (6) but there is considerable discordance between radiographic findings and clinical presentation that is not fully understood (7). MRI has shown some promises (8) but most orthopedic surgeons still rely on standard weight bearing radiographs. Improving the reliability and reproducibility of these interpretations could thus be hugely beneficial.

New technological advancement and recent progress of medical image analysis using deep learning (DL), a form of artificial intelligence (AI) has shown promising ways to detect knee OA (9). Traditional machine learning (ML) has often put a lot of effort into extracting features before training the algorithms, but in DL, we allow the algorithm to learn the features on its own from the data. This has turned out to be a hugely successful approach and opens up new ways to nonexperts in the field of ML to implement their own research and applications, such as medical image analysis, by shifting feature engineering from humans to computers (9).

There is to our knowledge only a few published studies applying DL to knee OA classification but many of these are applied to pre-processed, highly optimized images. Our aim was to see if we could train a neural network with an entire image series and not excluding common visual disturbances such as implants, casts and other pathologies.

Methods

Study design and setting

This study is a validation study of a diagnostic method based on retrospectively collected radiographic examinations. These examinations were analyzed by a neural network for both presence and severity of knee OA using the Kellgren & Lawrence OA classification system (KL) (10).

Data selection

We extracted all plain radiographic series taken between 2002 and 2016 from Danderyd University Hospital from the hospital's Picture Archiving and Communication System (PACS). Images along with corresponding radiologist reports were anonymized. For the purpose of this study, a random subset of image series of the knee were selected. In the test-set, the selection was biased towards OA and certain KL grade subtypes to reduce the risk of non-OA cases to dominating the evaluation data.

Inclusion & exclusion criteria

Projections included were not standardized. Trauma protocols as well as non-trauma protocols were included. Diaphyseal femur and tibia/fibula protocols were included as these display the knee joint

although not in the center of image. To ensure that the same case was not used twice, we excluded any examination within 90 days of the previously examination. We excluded 0.6 % of cases due to poor image quality as these could preclude classification.

Method of classification

In this method of machine learning the neural network identifies patterns in images. The network is fed both the input (the radiographic images) and the information of expected output label (classifications of OA grade) in order to establish a connection between the features of the different stages of knee OA, (e.g. possible osteophytes, joint space narrowing, sclerosis of subchondral bone) and the corresponding category (11).

Prior to being fed to the network the images along with radiologist's reports were uploaded to a custom-built platform to be labelled according to the KL grading scale by members of the research team (SO, AL, MG, EA). The KL grading scale was chosen as it is widely used for knee OA classification (12). The research team also evaluated any potential OA features in the patellofemoral joint when laterality radiographs were present. The lack of recognition of patellofemoral OA as a distinct or contributory factor has been a criticism of the KL grading scale (12). We also created custom output categories such as medial/lateral OA as it is interesting to see how well the network can discern these qualities on its own.

Data sets

We split the data into three sets: training, validation and test set. The same patient seeking and receiving an x-ray of the knee joint on multiple occasions with a > 90-day separation could be included multiple times in the same set, but there was no patient overlap between the training, validation and test sets.

The test set consisted of 300 cases classified by two senior orthopedic surgeons working independently. Any disagreement was dealt with by a joint reevaluation of the cases in question until a consensus between the two surgeons was reached. The test set then served as the ground truth that the final network was tested against.

During training, two sets of images were used. The training set which the network learned from and a validation set for evaluating performance and tweaking network parameters. The validation set was prepared in the same way as the test set but by SO and AL, two 4th year medical students. The training set was labeled only once by either SO or AL. If images were of bad quality or difficult to label, the students marked for revisit and were validated by MG. Initially, images were randomly selected for classification and fed to the network i.e. passive learning. As the learning progressed cases were selected based on the networks output: 1) initially cases with high probability of a class were selected to populate each category, and then 2) cases where the network is the most uncertain to define the border was used i.e. active learning (13).

Neural network setup

Network used was a of a ResNet type, convolutional neural network, that consisted in total of a 35-layer architecture with batch normalization for each convolutional layer and adaptive max pool (see Table 1 for structure). We randomly initialized the network and trained using stochastic gradient descent. During training we alternated between knee labels and other previously gathered fracture classification tasks (16,785 exams from other classification tasks (14) where each task shared the core network.

We initially trained without noise for 100 epochs with an initial learning rate of 0.025. After the initial training we re-set the learning rate to 0.01 and trained for another 50 epochs with a combination of white noise (5%) and random erasing 3 blocks per image of 10x10 pixels.

Type	Blocks	Kernel Size	Filters	Group
ResNet block	1	3x3	64	Image
ResNet block	1	3x3	64	Image
ResNet block	6	3x3	64	Core
ResNet block	4	3x3	128	Core
ResNet block	2	3x3	256	Core
ResNet block	2	3x3	512	Core
Image max	1	-	-	Pool
Convolutional	1	1x1	72	Classification
Fully connected	1	1x1	4	Classification
Fully connected	1	1x1	4	Classification

The images were additionally augmented with 2 jitters and separately processed up until a max pool merged the features into per image or exam depending on the type of outcome. In addition to the AO classification outputs that were pooled at the per exam level we had image view (i.e. AP, lateral, Oblique).

Input images

The network was presented with all available radiographs in each series. Each DICOM format radiograph was automatically cropped to the active image area, i.e. any black border was removed, and the image was reduced to a maximum of 256 pixels. Image dimensions were retained by padding the rectangular image to a square format of 256 x 256 pixels. During training we jittered the images by cropping and rotating.

Outcome measures/ statistical analysis

Network performance was measured using area under curve (AUC) as primary outcome measure and sensitivity, specificity and Youden J as secondary outcome measures. Proportion of correct graded OA severity was estimated using AUC - the area under a receiver operating curve (ROC) - which is a plot of true positive rate against the false positive rate. An AUC value of 1 signifies prediction that is always correct and a value of 0.5 is no better than random chance at predicting an outcome. There is no exact guide for how to interpret AUC values, but in general an AUC of 0.7-0.8 is considered acceptable, 0.8-0.9 is considered good or very good and ≥ 0.9 is considered outstanding (15, 16). Youden Index (J) is a value also used in conjunction with the ROC curve, it is a summary of sensitivity and specificity. It has a range of 0 to 1 and is defined as (17).

We also present confusion matrices to allow for good visualization of the algorithm's performance when the true values are known. The network was implemented and trained using PyTorch (v. 1.4). Statistical analysis was performed using R (4.0.0).

Results

Outcome Data

We included 5700 cases in the training set, 403 cases in the validation set and 300 cases in the test set. There was no patient overlap between the test set and training datasets. The most common KL grade in the training set and the test set was KL grade 0 and KL grade 3 respectively. In the test set, KL grade 3 was the most common type, closely followed by 4 and 0. Implants were used as a major visual disturbance to put more stress on the DL network. 11% of cases in the training set and 20% of cases in the test set had some type of a visible implant (Table 2).

Table 2: Distributions between KL grades and implants in different data sets.								
	Train				Test			
	Yes		No		Yes		No	
	n	(%)	n	(%)	n	(%)	n	(%)
Osteoarthritis								
Kellgren-Lawrence								
0	2,848	(50)	2,852	(50)	67	(22)	233	(78)
1	1,218	(21)	4,482	(79)	23	(8)	277	(92)
2	652	(11)	5,048	(89)	24	(8)	276	(92)
3	597	(10)	5,103	(90)	116	(39)	184	(61)
4	380	(7)	5,320	(93)	70	(23)	230	(77)
Other								
Medial	487	(9)	5,213	(91)	120	(40)	180	(60)
Lateral	175	(3)	5,525	(97)	34	(11)	266	(89)
Patella	36	(1)	2,486	(44)	3	(1)	297	(99)
Implant								
Implant	611	(11)	5,089	(89)	61	(20)	239	(80)
TKR	226	(4)	5,474	(96)	39	(13)	261	(87)
Unicompartmental	44	(1)	5,656	(99)	7	(2)	293	(98)
Plate	111	(2)	5,589	(98)	7	(2)	293	(98)
IM-nail	99	(2)	5,601	(98)	0	(0)	300	(100)
IM-nail femur	68	(1)	5,632	(99)				
IM-nail tibia	14	(0)	5,686	(100)	0	(0)	300	(100)
Cerclage	55	(1)	5,645	(99)	0	(0)	300	(100)
K-wires	21	(0)	5,679	(100)	0	(0)	300	(100)
Staple	8	(0)	2,514	(44)	4	(1)	296	(99)
X-fix	3	(0)	5,697	(100)	0	(0)	300	(100)
Screws	57	(1)	5,643	(99)	2	(1)	298	(99)

Table 2: Distributions between KL grades and implants in different data sets.								
	Train				Test			
	Yes		No		Yes		No	
	n	(%)	n	(%)	n	(%)	n	(%)
X-ligament	36	(1)	5,664	(99)	5	(2)	295	(98)

Abbreviations: TKA – Total knee replacement, IM – intramedullary, X-fix – external fixation, X- ligament – signs of ACL reconstruction.

Network results

All five KL grades displayed AUC > 0.80 with highest AUC for KL 0 with an AUC of 0.97, sensitivity and specificity of 97% and 88% respectively (Table 3). KL grade 2 had the lowest single performance with a sensitivity of 92%, specificity of 61% and an AUC of 0.80. When merging KL grades together generating larger groups, the network performed with AUCs of > 0.95 for all but the mid-ranged KL grade (KL 1, 2 and 3) which displayed an AUC of 0.82. The network performed excellent in differencing between medial and lateral OA (Table 4).

Table 3: Network performance on outcome measures. Kellgren & Lawrence grades are displayed separately and merged together.

	Cases (n=300)	Sensitivity (%)	Specificity (%)	Youden's J	AUC (95% CI)
Kellgren-Lawrence					
0	67	97	88	0.85	0.97 (0.93 to 0.99)
1	23	96	75	0.70	0.88 (0.83 to 0.92)
2	24	92	61	0.53	0.80 (0.73 to 0.86)
3	116	92	71	0.63	0.87 (0.83 to 0.90)
4	70	84	78	0.63	0.87 (0.83 to 0.91)
Grouped Kellgren-Lawrence					
0 to 1	90	88	95	0.83	0.96 (0.94 to 0.98)
0 to 2	114	83	97	0.81	0.97 (0.95 to 0.98)
1 to 3	163	80	74	0.53	0.82 (0.77 to 0.87)
2 to 4	210	97	83	0.80	0.96 (0.94 to 0.98)
3 to 4	186	96	88	0.84	0.97 (0.95 to 0.99)

Table 4: Network performance on differencing between medial or lateral OA					
	Cases (n=300)	Sensitivity (%)	Specificity (%)	Youden's J	AUC (95% CI)
Lateral	34	100	89	0.89	0.97 (0.95 to 0.98)
Medial	120	91	88	0.79	0.94 (0.91 to 0.97)
Osteonecrosis	13	85	50	0.34	0.65 (0.54 to 0.75)
Patella	3	100	99	0.99	1.00 (0.99 to 1.00)

The confusion matrices between the true (senior orthopedic consultants) and predicted label (AI-network) demonstrates the DL networks ability to classify the test set to that of an orthopedic expert. As indicated

by the AUC values, the network had most difficulties deciding category for KL grade 1 and 2.

Network decision analysis

We sampled cases for analysis where the network was most certain of a prediction, whether correct or incorrect. Examples of various KL grades are shown below. (figures 2a-d). Heatmaps visualizing areas in an image the network focuses on are shown as colored dots. There were no clear discernable trends among these cases as to what made the network falter or succeed.

When analyzing the network activity, we could find that many of the missed labels were subjective, e.g. in figure 2a we find that the network suggests a class of 1-2 while the true class is 0. The heatmap activity is centered around the implant and could be remaining indicators from the already operated medial arthrosis.

Discussion

This study further demonstrates the potential of implementing AI- networks to aid in classifying knee OA severity. We believe this study, in accordance with other studies in the field, demonstrates DL potential in OA classification. Our network yielded an overall high AUC of >0.87 for all KL grades except KL grade 2 and a mean AUC of 0.92. When merging adjacent KL grades together, all but one group showed near perfect results with $AUC > 0.95$. This further display that the network predictions is often close to the ground truth (as established by two orthopedic surgeons).

Network performance was generally excellent for all KL grades except for KL grade 2 which had an AUC of 0.8. It is expected that images in the middle of the spectrum will be more difficult to assign proper category both for humans and DL networks and these mid-categories have also more vague severity definitions. This is also demonstrated as we merged adjacent KL grades together, all but mid-ranged KL grade (KL 1, 2 and 3) showed near perfect results. We can also conclude that the AI- network was able to locate medial / lateral OA in the test set with high precision.

In contrast to similar studies by Tiulpin (18) and Norman (19), we let our network learn from entire image series without any preprocessing. All images used in the training set was taken from a general setting where radiographs can differ greatly in quality and beam angels. Both Tiulpin and Norman used a region-of-interest attention zone around the knee joint to alleviate their algorithms to learn more features than needed whereas we trained out network on whole radiographs and with the knee joint in different positions. Tiulpin et al furthermore exclude images with implants to avoid disturbance in data distribution. In contrast, we let our network learn from the whole image and to locate the knee joint by itself along with a high proportion of implants. Our algorithm also shows the possibility to train and test a

DL network on a relatively small number of radiographs and still attain AUC close to that of the aforementioned studies.

Limitations

One limitation is the lack of different DL architecture tested. There is a vast amount of DL architectures in the DL domain and possibly some of these could have improved our results. It was although not our aim to find the most superior DL architecture in this study. Furthermore, the network can only be as good as the grading system it learns from. KL grading scale is the most used system for classifying knee OA. It is however not a perfect grading scale with vague mid categories descriptions (12). This makes ground truth being subjective to user preference, something that the DL network is not able to cope with. Prior studies (10, 20, 21) have reported that KL suffers from ambiguity with interobserver reliability ranging from 0.5- 0.9. Culvenor et al. (22) made a study that compared knee OA using two different grading scales, KL and Osteoarthritis Research Society International (OARSI). They concluded that tibiofemoral OA was twice as common using OARSI grading scale compared to the KL system. There is a possibility we could have increased our accuracy of early osteoarthritis (grade 1-2) by using the OARSI system instead of KL.

While the radiographs used were collected from over a decade long period with a large sample of patients, our selection was limited in that the data source is a single hospital in Stockholm. From a generalization point of view, this study had disadvantages since all images were taken from PACS at Danderyd University Hospital with a mainly Caucasian population. The neural network would potentially present different results if tested on radiographs outside Sweden where ethnicity and joint alterations may differ.

Clinical applications and future studies

As technology advances, so could expectation on classification accuracy. Further into the future a possible task could be having the network calculate success of different treatment strategies given features of the OA. It is however important to understand that OA is by nature a progressive disease with no clear boundary between KL grades. This consequently makes it impossible for a network to perform a perfect result. Correct application of a classification system for OA can nonetheless point towards the degree of severity and its progress alongside the clinical assessment of the patients, aiding physicians in evaluating the necessary treatment plans.

In a future study, it would be interesting to include patient's symptoms and clinical signs in addition to radiographic findings in a DL network. This could later be used to analyze different orthopedic clinics in how keen they are to operate TKR-surgery on different KL grades. As MRI becomes more available and cheaper, a shift from weight-bearing plain radiographs to MRI would make studies between different modalities and DL noteworthy as well.

Conclusion

We found that we could teach a neural network classify knee OA severity and laterality using the KL grading scale without cleaning the input data from major visual disturbances such as implants and other pathologies.

Declarations

Availability of data and materials heading: The test dataset used during evaluation phase of the current study available from <https://datasets.aida.medtech4health.se/>.

Competing Interest: MG and AS are co-founders and shareholders in DeepMed AB.

Ethics Approval and Consent to Participate: The research was approved by the Swedish Ethical Review Authority (dnr: 2014/453-31)

Author's Contribution: SO (Conceptualization; Data curation; Formal analysis; Writing – original draft; Writing – review & editing). AL (Conceptualization; Data curation; Writing – review & editing). AL (Data curation; Writing – review & editing). ASR (Methodology; Software; Writing – review & editing). MG (Conceptualization; Data curation; Formal analysis; Funding acquisition; Investigation; Project administration; Resources; Software; Supervision; Validation; Visualization; Writing – review & editing). All authors have read and approved the manuscript.

Funding: This project was supported by grants provided by Region Stockholm (ALF project) that have enabled both research time and computational resources for MG & ASR. The grants are part of a larger project and are not linked specifically to the current manuscript.

Acknowledgments: We would like to thank prof. Olof Sköldenberg for his help with classifying images in the test-set.

References

1. Sasek C. An update on primary care management of knee osteoarthritis. JAAPA : official journal of the American Academy of Physician Assistants. 2015;28(1):37-43.
2. Nemes S, Rolfson O, W-Dahl A, Garellick G, Sundberg M, Kärrholm J, et al. Historical view and future demand for knee arthroplasty in Sweden. Acta orthopaedica. 2015;86(4):426-31.
3. Murphy L, Schwartz TA, Helmick CG, Renner JB, Tudor G, Koch G, et al. Lifetime risk of symptomatic knee osteoarthritis. Arthritis & rheumatism. 2008;59(9):1207-13.
4. Kurtz S, Ong K, Lau E, Mowat F, Halpern M. Projections of primary and revision hip and knee arthroplasty in the United States from 2005 to 2030. Journal of bone and joint surgery. 2007;89(4):780-5.

5. Gwilym S, Pollard T, Carr A. Understanding pain in osteoarthritis. *The Journal of bone and joint surgery British volume*. 2008;90(3):280-7.
6. Ho-Pham LT, Lai TQ, Mai LD, Doan MC, Pham HN, Nguyen TV. Prevalence of radiographic osteoarthritis of the knee and its relationship to self-reported pain. *PloS one*. 2014;9(4):e94563.
7. Hannan MT, Felson DT, Pincus T. Analysis of the discordance between radiographic changes and knee pain in osteoarthritis of the knee. *The Journal of rheumatology*. 2000;27(6):1513-7.
8. Barr AJ, Campbell TM, Hopkinson D, Kingsbury SR, Bowes MA, Conaghan PG. A systematic review of the relationship between subchondral bone features, pain and structural pathology in peripheral joint osteoarthritis. *Arthritis research & therapy*. 2015;17(1):228.
9. Shen D, Wu G, Suk H-I. Deep Learning in Medical Image Analysis. *Annual review of biomedical engineering*. 2017;19:221-48.
10. Kellgren JH, Lawrence JS. Radiological assessment of osteo-arthritis. *Annals of the rheumatic diseases*. 1957;16(4):494-502.
11. Erickson BJ, Korfiatis P, Akkus Z, Kline TL. Machine Learning for Medical Imaging. *Radiographics*. 2017;37(2):505-15.
12. Kohn MD, Sassoon AA, Fernando ND. Classifications in Brief: Kellgren-Lawrence Classification of Osteoarthritis. *Clinical orthopaedics and related research*. 2016;474(8):1886-93.
13. Smailagic A, Costa P, Noh HY, Walawalkar D, Khandelwal K, Galdran A, et al., editors. MedAL: Accurate and robust deep active learning for medical image analysis. 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA); 2018: IEEE.
14. Olczak J, Emilson F, Razavian A, Antonsson T, Stark A, Gordon M. Ankle fracture classification using deep learning: automating detailed AO Foundation/Orthopedic Trauma Association (AO/OTA) 2018 malleolar fracture identification reaches a high degree of correct classification. *Acta Orthopaedica*. 2020:1-7.
15. Fangyu L, Hua H. Assessing the accuracy of diagnostic tests. *Shanghai archives of psychiatry*. 2018;30(3):207.
16. Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology*. 2010;5(9):1315-6.
17. Hajian-Tilaki K. Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. *Caspian journal of internal medicine*. 2013;4(2):627.
18. Tiulpin A, Thevenot J, Rahtu E, Lehenkari P, Saarakkala S. Automatic Knee Osteoarthritis Diagnosis from Plain Radiographs: A Deep Learning-Based Approach. *Sci Rep*. 2018;8(1):1727.
19. Norman B, Pedoia V, Noworolski A, Link TM, Majumdar S. Applying Densely Connected Convolutional Neural Networks for Staging Osteoarthritis Severity from Plain Radiographs. *Journal of digital imaging*. 2019;32(3):471-7.
20. Wright RW. Osteoarthritis Classification Scales: Interobserver Reliability and Arthroscopic Correlation. *Journal of bone and joint surgery*. 2014;96(14):1145-51.

21. Gossec L, Jordan JM, Mazzuca SA, Lam MA, Suarez-Almazor ME, Renner JB, et al. Comparative evaluation of three semi-quantitative radiographic grading techniques for knee osteoarthritis in terms of validity and reproducibility in 1759 X-rays: report of the OARSI-OMERACT task force. Osteoarthritis and cartilage. 2008;16(7):742-8.

22. Culvenor AG, Engen CN, Øiestad BE, Engebretsen L, Risberg MA. Defining the presence of radiographic knee osteoarthritis: a comparison between the Kellgren and Lawrence system and OARSI atlas criteria. Knee surgery, sports traumatology, arthroscopy : official journal of the ESSKA. 2015;23(12):3532-9.

Figures

Figure 1

Confusion matrices between true label and predicted label in the test set. The numbers on X and Y- axis represents KL grades 0-4. The diagonal represents the accurate prediction the labelers and the AI-network.

Figure 2

Incorrectly classified KL grade 0. Incorrectly classified cases where the network failed to detect OA KL 0. The integrated gradient heatmap shows features that contribute to the case being incorrectly classified as red dots while blue dots indicate features weighing against that output. Correctly classified KL grade 1. Correctly classified cases where the network failed to detect OA KL 0. The integrated gradient heatmap shows features that contribute to the case being incorrectly classified as red dots while blue dots indicate features weighing against that output. Incorrectly classified KL grade 4. Incorrectly classified cases where the network wrongly detected as OA KL 4. The integrated gradient heatmap shows features that contribute to the case being incorrectly classified as red dots while blue dots indicate features weighing against that output. Correctly classified KL grade 4: Correctly classified KL grade 1 by the AI network. The integrated gradient heatmap shows features that contribute to the case being correctly classified as red dots while blue dots indicate features weighing against that output.